PROBABILISTIC APPROACH FOR EVALUATING METABOLITE SAMPLE INTEGRITY

BARRY M. SLAFF, SHANE T. JENSEN, AND AALIM M. WELJIE

Abstract. The success of metabolomics studies depends upon the "fitness" of each biological sample used for analysis: it is critical that metabolite levels reported for a biological sample represent an accurate snapshot of the studied organism's metabolite profile at time of sample collection. Numerous factors may compromise metabolite sample fitness, including chemical and biological factors which intervene during sample collection, handling, storage, and preparation for analysis. We propose a probabilistic model for the quantitative assessment of metabolite sample fitness. Collection and processing of nuclear magnetic resonance (NMR) and ultra-performance liquid chromatography (UPLC-MS) metabolomics data is discussed. Feature selection methods utilized for multivariate data analysis are briefly reviewed, including feature clustering and computation of latent vectors using spectral methods. We propose that the time-course of metabolite changes in samples stored at different temperatures may be utilized to identify changing-metabolite-to-stable-metabolite ratios as markers of sample fitness. Tolerance intervals may be computed to characterize these ratios among fresh samples. In order to discover additional structure in the data relevant to sample fitness, we propose using data labeled according to these ratios to train a Dirichlet process mixture model (DPMM) for assessing sample fitness. DPMMs are highly intuitive since they model the metabolite levels in a sample as arising from a combination of processes including, e.g., normal biological processes and degradation- or contamination-inducing processes. The outputs of a DPMM are probabilities that a sample is associated with a given process, and these probabilities may be incorporated into a final classifier for sample fitness.

1. Introduction

Quantitative analysis of metabolite levels in biofluids and tissues has become a fruitful approach in medical and translational research. The success of such studies depends upon the "fitness" of each biological sample used for analysis. Specifically, it is critical that metabolite levels reported for a biological sample represent an accurate snapshot of the studied organism's metabolite profile at time of sample collection. Between sample collection and final analysis, numerous factors may compromise metabolite sample fitness, including chemical (e.g. thermodynamic) and biological (e.g. bacterial) factors which intervene during sample collection, handling, storage, and preparation for analysis. We propose a probabilistic model for the quantitative assessment of metabolite sample fitness. The proposed model may be implemented as a computational tool which uses the measured metabolite profile from a biological sample to estimate the sample's fitness for inclusion in metabolomics analyses.

In the present work, we present an approach for developing a probabilistic model of metabolite sample fitness. In Section 2 we discuss the proposed analytical methods for collecting metabolite sample data. The proposed analytical platforms are nuclear magnetic resonance (NMR) spectroscopy with a targeted profiling approach for quantitation together with ultra-performance liquid chromatography coupled to mass spectrometry (UPLC-MS). These are established methods for the acquisition of quantitative metabolite data from biofluid and tissue samples. Additionally, we discuss the critical data processing steps of normalization, centering, and scaling. Finally, we discuss approaches to feature selection for the final model-building process.

In Section 3.1 we discuss the principle of metabolite sample fitness. Critical to the modeling process is a quantitative definition of sample fitness: how should fitness be assessed? Existing studies suggest the possibility of identifying metabolites that change significantly in concentration ("changing metabolites") and metabolites which remain relatively stable ("stable metabolites") during sample storage at different temperatures. We propose that at least one changing-metabolite-to-stable-metabolite ratio might be identified for each biological matrix, and tolerance intervals will be computed to characterize the range of ratios among fresh samples. Therefore, a sample can be deemed fit or unfit based on whether its changing-metabolite-to-stable-metabolite ratios fall within the computed tolerance intervals for fresh samples. While this method offers a straightforward computation, it does not capture all the structure available in the data for differentiating fresh and degraded samples. Therefore, we propose that data classified using tolerance intervals may be used to train a probabilistic model capable of capturing additional structure in the data.

In Section 3.2, we propose approaches for modeling metabolite sample fitness based on the principle discussed in Section 3.1. Key to our approach is development of a model which avoids incorrect parametric assumptions. Our main modeling approach is development of a Dirichlet process mixture model (DPMM) for each biological matrix for which we wish to assess sample fitness. DPMMs are highly intuitive since they model the metabolite levels in a sample as arising from a combination of processes including, e.g., normal biological processes and degradation- or contamination-inducing processes. The outputs of a DPMM are probabilities that a sample is associated with a given process, and these probabilities will be incorporated into a final classifier for sample fitness. We also consider the use of conceptually simple non-parametric methods such as k-Nearest-Neighbor and kernel regression.

2. Model Inputs: Data Acquisition and Processing

Nuclear Magnetic Resonance (NMR) Spectroscopy and Ultra-Performance Liquid Chromatography - Mass Spectrometry (UPLC-MS) are state-of-the-art analytical platforms of the acquisition of quantitative metabolite data from urine, plasma, serum, tissue, and other biological matrices [1–5]. Quantitative data obtained using NMR and UPLC-MS will be utilized to model and assess metabolite sample integrity.

2.1. **Data Acquisition.**

Experimental design and data acquisition in metabolomics must avoid contamination of the data with systemic errors and variances that can compromise analyses [6]. Targeted profiling [7,8] with NMR spectroscopy produces quantitative metabolite concentrations which are reproducible within [9, 10] and between [11, 12] labs, with more variation when NMR probes and experimental parameters are not consistent [13]. A Design of Experiments (DoE) approach together with UPLC-MS yields reproducible metabolite data in quantitative and non-quantitative approaches [14–16]. Appropriate measures will be taken in the experimental design to avoid biases arising from test subject selection and temporal factors (e.g. time of day of sample collection). Sample preparation and storage procedures will be tightly controlled apart from those varied deliberately as part of the experimental procedure for inducing sample degradation. Our investigation of metabolite sample integrity will inform existing domain knowledge regarding proper sample preparation and storage for metabolomics studies [17–23] and biobanks [24–26].

2.2. **Data Processing.** Normalization, centering, and scaling are essential steps for metabolomics data processing. Normalization is a critical step for purposes of comparing spectra acquired from samples with different levels of dilution [27–29]. This concern is particularly critical in the case of urine samples [29, 30]. The simplest normalization method is total integral normalization, which assumes that all spectral peaks scale with sample dilution. We also consider probability quotient normalization [31] (PQN), also called median-fold change normalization [32] (MFC), which scales the peaks in each spectrum by one factor per spectrum so that the median fold-change between the peaks in each spectrum and corresponding peaks in a reference spectrum is 1. In contrast to total spectral normalization, PQN/MFC assumes that most rather than all spectral peaks scale with sample dilution. We also consider an additional normalization step which would minimize the distance from the sample vector to a modeled probability distribution. The metric used to evaluate nearness might be Euclidean distance as a default, or for example Mahalanobis distance [33] if the clusters are modeled as multivariate Gaussians. In the case of Gaussian clusters with diagonal covariance, this computation is analytically simple [34] and it is analytically or numerically computable in other cases. Total integral normalization (TIN) and PQN/MFC are widely used and have been studied comparatively in the context of both NMR [27] and LC-MS [28] metabolomics data. It has been shown that use of TIN or PQN/MFC improves the results of comparing spectra relative to no normalization, while the optimal method varies between contexts [27–29].

Prior to model-building with training data and classification with new data, the data may be mean-centered and scaled feature-by-feature. Scaling assumes that the features with the most variance are not necessarily the features with the most predictive value, since relatively abundant features tend to have greater variance. Several scaling methods widely used in metabolomics studies [35, 36] include auto-scaling (each feature in the training data is scaled to have variance 1), Pareto scaling (each feature scaled so that its variance is the square-root of its initial variance), Variable stability (VAST) scaling [37] (features with smaller coefficient of variation are given more weight), and range scaling (each feature is scaled by its full range). The optimal scaling approach has been found to be highly context-dependent, and the results of modeling depend significantly upon the scaling method utilized [35, 36].

2.3. **Feature Selection.** It is often advantageous in data analysis contexts to utilize a subset of the acquired features or generate new features for the final modeling task. For example, eliminating irrelevant or noisy features can improve the predictive performance of any final model. Additionally, generating new features which are highly relevant to the prediction can improve model performance.

With respect to choosing subsets of acquired features for modeling, we consider the following:

(1) Since the consistently-detectable metabolites are known to differ across matrices for NMR [1] and LC-MS [3], features should be selected on a per-matrix basis, i.e. one feature set for human urine, one feature set for human serum, etc. The full panel of reliably-detectable metabolites will be profiled initially.

(2) A Design-of-Experiment approach [14, 38, 39] (DoE) may be used to experimentally narrow the matrix-specific feature lists. For example, metabolites common to the same molecular pathways may exhibit high co-linearity, which would be detectable with a DoE approach.

(3) In the case of a parametrized clustering approach, it may be useful to include only features which satisfy certain parametric constraints. For example, in a model which constructs

multivariate Gaussian clusters, it may be interesting the maximize the Gaussian character of the joint distributions through feature selection. Multivariate Gaussian character can be assessed using an R package such as MVN [40]. This may be achieved by selecting only features (possibly after transformation) with univariate Gaussian character, which can be assessed with a univariate test for normality such as Lilliefors [6, 41, 42] with an R package such as nortest [43].

It may be desirable to combine the initial metabolite features into new features for purposes of modeling metabolite sample fitness. Numerous feature selection methods have been utilized in omics studies, in particular in domains for which studies typically include many more features than samples [44,45]. Possible feature-combination approaches for modeling metabolite sample fitness include:

(1) Components from unsupervised spectral methods: utilizing principal component analysis (PCA) vectors (equivalently, singular vectors and values) or independent component analysis (ICA) vectors as features [46].

(2) Latent vectors: utilizing latent vectors from partial least squares regression [47] (PLS), canonical correlation analysis [48] (CCA), or related spectral methods such as orthogonal PLS [49] (O-PLS) and O-PLS with discriminant analysis [50] for classification (O-PLS-DA). The orthogonalized methods remove systemic variation in the data unrelated to the response variables to improve interpretability. O-2-PLS [51] additionally yields two-way information about covariation and predictiveness between the observed and response variables. For these methods, the response variables could be, for example indicators of sample degradation such as time of sample exposure to non-freezing storage temperature. These methods have been used widely for metabolomics data analysis [21, 52, 53].

(3) Feature clustering: The method of shrunken centroids [54], which originated as a feature selection method in genomics, has been applied in a metabolomics context [55,56] for choosing a subset of highly representative features. Other clustering approaches involve using mutual information [57] or using a graph-theoretic approach [58] to identify clusters and choose representative metabolite features.

(4) Multiple-testing framework for discovering significant features: kernelized support vector machines [59], k-nearest-neighbor [60], and classificaion trees [61] have been used together in a multiple testing framework to identify individual metabolite features [62]. This approach is not widely used in the metabolomics literature but is attractive particularly since the false discovery rate can be controlled via the multiple testing framework.

3. Modeling Sample Fitness

We wish to distingish fit samples from unfit samples. In principle, a sample is fit for analysis if its measurable metabolite levels at time of analysis are very similar to its measurable metabolite levels at time of collection. According to this principle, a sample is fit if at time of measurement, it accurately captures a collection-time snapshot of the studied organism's metabolite profile. The original metabolite levels change over time due to intervention from chemical and biological factors during sample handling, storage, and preparation for analysis. Therefore, our central problems are the following:

(1) Quantify the degree of change in metabolite levels after collection for which the sample should no longer be considered "fit" for analysis. We follow the recommendation of Fraser et al [63], now widely adopted for quality control reporting in clinical medicine [64], and define three levels of sample fitness: optimal, desirable, and minimal.

(2) Construct a probabilistic model of metabolite sample fitness so that a sample can be accurately categorized as fit (optimal, desirable, or minimal) or unfit based upon its reported metabolite levels. The model will be trained on data for which we have followed the time-course of metabolite level changes from the absolutely fresh state to various states of degradation. The result of (1) will be used to label each training data point as fit (optimal, desirable, or minimal) or unfit. The trained probabilistic model will be used to predict the fitness or non-fitness of samples for which we have only one measurement of metabolite levels.

Problem (1) is the subject of Section 3.1 and (2) is the subject of Section 3.2.

3.1. **Principle of Metabolite Sample Fitness.** Recent studies identify urine, blood, and plasma metabolites that change concentration significantly over hours and days in response to storage at above-freezing temperatures [18–20, 23]. This degradation process transforms a sample from a state of fitness to unfitness. We hypothesize that ratios between changing metabolites and stable metabolites during storage can identify fresh vs. degraded samples. For this purpose we propose the use of tolerance intervals [65, 66] for characterizing changing-to-stable metabolite ratios in optimally-fit, desirably-fit, and minimally-fit samples. For each biological matrix, at least one changing/stable metabolite pair should be identified from the available data. We propose the following taxonomy:

(1) An "optimally" fit sample is one which falls inside a tolerance interval containing 80% of the fresh-sample ratios with 95% confidence (i.e., .80-content, .95-coverage TI).
(2) A "desirably" fit sample is one which falls inside a tolerance interval containing 95% of the fresh-sample ratios with 95% confidence (i.e., .95-content, .95-coverage TI) and which is not "optimally" fit.
(3) A "minimally" fit sample is one which falls inside a tolerance interval containing 99% of the fresh-sample ratios with 95% confidence (i.e., .99-content, .95-coverage TI) and which is not "desirably" fit.

Tolerance intervals may be computed for data arising from approximately normal [67] or non-normal distributions [68]. In the case of approximately normal distributions, the two-sided p-content, γ-coverage tolerance interval is

$$[\bar{Y} - s \cdot k, \bar{Y} + s \cdot k],$$

where $\bar{Y}$ is the sample mean, s is the sample standard deviation, and k is computed as:

$$k = \sqrt{\frac{\nu \left(1 + \frac{1}{N}\right) z^2_{(1-p)/2}}{\chi^2_{1-\gamma,\nu}}}, \tag{1}$$

where $\chi^2_{1-\gamma,\nu}$ is the critical value of the chi-square distribution with degrees of freedom ν (usually $\nu = N - 1$) that is exceeded with probability γ, and $z_{(1-p)/2}$ critical value of the normal distribution associated with cummulative probability (1-p)/2.

Changing-to-stable metabolite ratios offer a straightforward method for identifying fit and unfit samples. However, this method is not comprehensive: it may not identify all structure present in the available data for differentiating fresh and degraded samples. We therefore propose that changing-to-stable metabolite ratios may be used to train a probabilistic model which is capable of capturing additional structure in the data. We propose to utilize a Dirichlet process mixture model for this purpose (see Section 3.2.2) and also consider the use of k-Nearest-Neighbor and kernel regression methods (see Section 3.2.1).

3.2. **Modeling Methods.** We discuss two main approach types for modeling metabolite sample fitness. The first approach type includes k-nearest neighbors (kNN) and kernel regression, two conceptually simple approaches not widely utilized for classification in the metabolomics literature. They are an intuitive method for classifying fit- and unfit samples, since fit samples should be more similar to other fit samples than unfit samples. However, kNN and kernel regression do not identify important relationships between significant features and are highly suceptible to misclassification due to contributions from insignificant features. Hence we also present a second approach.

The second approach type is that of Dirichlet process mixture models (DPMM), a non-parametric Bayesian approach [69]. Despite the “non-parametric” descriptor, a key advantage of such models is not their absence of parameters but their flexible parametric form: the parametric form of the probability distributions is inferred along with the parameter values during model training. Characterization of metabolite sample fitness according to DPMM is highly intuitive because we regard sample feature measurements as arising from a combination of biological and chemical processes, including freshness (ideal sample collection and measurement) and degradation mechanisms including chemical (e.g. thermodynamic) and biological (e.g. bacterial) over varying lengths of time. Sample fitness may then be assessed according to the estimated probability of each process having generated a given sample.

3.2.1. *k-Nearest Neighbors and Kernel Regression.* The k-Nearest Neighbor (kNN) classifier [60, 70] is one of the simplest classifiers conceptually and has only one parameter, k. Given training samples and a new test sample, it classifies the new sample in the majority category of its k nearest training-sample neighbors. The nearness is computed according to a metric. Some metrics utilized for nearest neighbor classification include [71]:

(1) L_2 norm: normal Euclidean distance. The distance between two points is the square root of the sum of squared differences between the features.
(2) L_1 norm: The distance between two points is the sum of the absolute values of differences between the features.
(3) L_∞ norm: The distance between two points is the absolute value of the greatest difference between any two features.

The optimal k may be determined using cross-validation. An advantage to using kNN for modeling fit and unfit samples is that it makes no parametric assumptions about the probability distributions of the sample groups or about the way in which those groups separate (in contrast to, for example, a hyperplane-separation method). However, since all features are treated equally in computing the distance between two points, the method is prone to mis-classification without exceptionally careful feature selection. K-nearest-neighbor classifiers have been discussed in the metabolomics literature but are rarely utilized for classification, in part due to their relatively large computational expense [72].

An extension of kNN is *kernel regression* [71, 73, 74]. kNN assumes that when classifying a new sample, all training samples should receive equal consideration and the number of neighbors (k) should be the same for classifying any new sample. Kernel regression relaxes those assumptions by defining a kernel which gives different weight to each training sample in the classification. Generally speaking, a *kernel* is a similarity function which maps pairs of data points to a number; more similar ("nearer") data points should be mapped to larger numbers. One widely-used kernel is a Gaussian kernel, which gives stronger weight to nearer training samples and less weight to farther-away training samples according to an exponential fall-off. For Gaussian kernel regression, the prediction $\hat{y}$ for a new sample $\mathbf{x}$ is

$$\hat{y}(\mathbf{x}) = \underset{y \in Y}{\arg\max} \sum_{i=1}^{N} I(f(\mathbf{x}_i) = y) K(\mathbf{x}_i, \mathbf{x}), \tag{2}$$

where y is a class identifier, Y is the set of classes, N is the number of training samples, f maps a training sample to its class identifier, I is the indicator function (I=1 if the argument is true, I=0 otherwise), and K is the kernel function, in this case the Gaussian kernel [71]:

$$K(\mathbf{x}_i, \mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left[-\frac{1}{2}(\mathbf{x_i} - \mathbf{x})^T \Sigma^{-1} (\mathbf{x_i} - \mathbf{x})\right], \tag{3}$$

where k is the vector dimension of $\mathbf{x}$ (i.e. the number of features) and Σ is a k x k covariance matrix. Σ is a free parameter which may be determined by cross-validation or computed from the training data, e.g.

$$\Sigma = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}_\mathbf{i})(\mathbf{x}_i - \bar{\mathbf{x}}_\mathbf{i})^T, \tag{4}$$

where $\bar{\mathbf{x}}_\mathbf{i}$ is the average of the training samples.

Many kernels have their own parameters which must be determined by cross-validation; for example, the Gaussian kernel has a "spread" or "standard deviation" parameter σ governing how fast the exponential falls off. As the number of parameters increases, so does the risk for over-fitting the prediction model.

3.2.2. *Dirichlet Process Mixture Model.* We can conceptualize the sample fitness assessment problem as follows: we assume that samples are generated by a combination of distinct *processes* (i.e. probability distributions) and consider those processes to be hidden variables in a mixture model. For example, we might consider a human urine sample metabolite levels to result from processes such as freshness (ideal fresh urine sample collection), bacterial contamination (due to exposure to above-freezing storage temperatures) over varying lengths of time, and chemical interactions within the urine (e.g. breakdown of thermodynamically unstable compounds or slow reactions) over varying lengths of time. Our proposed experimental procedure includes collection and measurement of biofluid metabolite levels after variable-time exposure to a range of storage temperatures. We propose to model the resulting data using a Dirichlet process mixture model (DPMM). To assess the fitness of a new sample, the DPMM will estimate the maximum posterior probability that the sample arose from each process. In this way we obtain a quantitative estimate of the degree to which a sample is generated from, e.g., freshness vs. various degradation processes. To make a sample fitness determination, we consider the following possibilities:

(1) Based on the principle of sample fitness (see Section 3.1), the training samples may be used to establish a threshold for the probability of "fresh" process generation which constitutes a fit sample. Hence the final classifier predicts that the sample is fit or unfit based on the probability that the sample arose from the "fresh" process.

(2) However, more than one process may be associated with sample "freshness". Therefore it may be desirable to use the process probabilities computed by the DPMM as inputs to a final classifier for sample fitness. For this procedure, we will explore the use of classifiers such as kNN, O-PLS-DA [50], and hyperplane methods such as support vector machine [59] or soft independent modeling of class analogy [75] (SIMCA). Hence the final classifier assess sample fitness using the full vector of process probabilities computed by the DPMM.

The advantage of using a DPMM lies in its parametric flexibility: the priors for each process are defined over a space of functions, and the appropriate parametric representation of each process is inferred *together with* the parameter values. In contrast, for more strongly-parametrized approaches such as a Gaussian Mixture Model, the parametric representation of each process is assumed and only the parameter values are computed in model training.

References

[1] Keun H.C. and Athersuch T.J. Nuclear magnetic resonance (NMR)-based metabolomics. *Methods Mol Biol.*, 2011.

[2] Beckonert O. et al. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols 2 pp. 2692-703*, 2007.

[3] Zhou B. et al. LC-MS-based metabolomics. *Mol Biosyst.*, 2012.

[4] Want E et al. Global metabolic profiling of animal and human tissues via UPLC-MS. *Nature Protocols*, 2013.

[5] Benton H.P. et al. Intra- and Interlaboratory Reproducibility of Ultra Performance Liquid ChromatographyTime-of-Flight Mass Spectrometry for Urinary Metabolic Profiling. *Anal. Chem. 84 (5)*, 2012.

[6] Razali N.H. and Yap B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2011.

[7] Weljie A.M. et al. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem.*, 2006.

[8] Athersuch T.J. et al. Evaluation of 1H NMR metabolic profiling using biofluid mixture design. *Anal Chem.*, 2013.

[9] Weljie A.M. et al. Evaluating low-intensity unknown signals in quantitative proton NMR mixture analysis. *Anal. Chem.*, 2008.

[10] Tredwell G.D. et al. Between-Person Comparison of Metabolite Fitting for NMR-Based Quantitative Metabolomics. *Anal. Chem.*, 2011.

[11] Keun H.C. et al. Analytical reproducibility in (1)H NMR-based metabonomic urinalysis. *Chem. Res. Toxicol.*, 2002.

[12] Ravanbakhsh S. et al. Accurate, fully-automated NMR spectral profiling for metabolomics. *arXiv*, 2014.

[13] Lacy P. et al. Signal Intensities Derived from Different NMR Probes and Parameters Contribute to Variations in Quantification of Metabolites. *PLoS One*, 2014.

[14] Zheng H. et al. Time-Saving Design of Experiment Protocol for Optimization of LC-MS Data Processing in Metabolomic Approaches. *Anal. Chem.*, 2013.

[15] Xiao J.F. et al. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Analyt Chem.*, 2012.

[16] Oberacher H. et al. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *J Mass Spectrom.*, 2009.

[17] Dona A.C. et al. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem.*, 2014.

[18] Emwas A. et al. Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics*, 2014.

[19] Kamlage B. et al. Quality Markers Addressing Preanalytical Variations of Blood and Plasma Processing Identified by Broad and Targeted Metabolite Profiling. *Clinical Chemistry*, 2014.

[20] Yin P. et al. Preanalytical Aspects and Sample Quality Assessment in Metabolomics Studies of Human Blood. *Clinical Chemistry*, 2013.

[21] Korman A. et al. Statistical methods in metabolomics. *Methods Mol Biol.*, 2012.

[22] Bruce S.J. et al. Investigation of Human Blood Plasma Sample Preparation for Performing Metabolomics Using Ultrahigh Performance Liquid Chromatography/Mass Spectrometry. *Anal. Chem.*, 2009.

[23] Saude E.J. and Sykes B.D. Urine stability for metabolomic studies: effects of preparation and storage. *Metabolomics*, 2007.

[24] Elliot P. and Peakman T.C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.*, 2008.

[25] Dunn W.B. et al. A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *Int J Epidemiol.*, 2008.

[26] Hebels D.G.A.J. et al. Performance in Omics Analyses of Blood Samples in Long-Term Storage: Opportunities for the Exploitation of Existing Biobanks in Environmental Health Research. *Environ Health Perspect*, 2013.

[27] Giraudeau P., Tea I., Remaud G., and Akoka S. Reference and normalization methods: Essential tools for the intercomparison of NMR spectra. *Journal of Pharmaceutical and Biomedical Analysis*, 2014.

[28] Ejigu B.A. et al. Evaluation of Normalization Methods to Pave the Way Towards Large-Scale LC-MS-Based Metabolomics Profiling Experiments. *OMICS*, 2013.

[29] Wang B., A. Goodpaster, and Kennedy M. Coefficient of variation, signal-to-noise ratio, and effects of normalization in validation of biomarkers from NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems*, 2013.

[30] Slaff B., Sengupta A., and Weljie A.M. (In Press.) NMR Spectroscopy of Urine. In Keun H.C., editor, *NMR-based Metabolomics.* 2014.

[31] Dieterle et al. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.*, 2006.

[32] Veselkov KA et al. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal Chem.*, 2011.

[33] De Maesschalck R. et al. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 2005.

[34] Geitz. Vector Geometry for Computer Graphics. *https://www.cs.oberlin.edu/ bob/cs357.08/VectorGeometry/VectorGeometry.pdf.*

[35] van den Berg R. A. et al. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 2006.

[36] Gromsky P.S. et al. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics*, 2015.

[37] Keun H.C. et al. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Analytica Chimica Acta*, 2003.

[38] Riter L. et al. Statistical design of experiments as a tool in mass spectrometry. *J Mass Spectrom*, 2005.

[39] Eliasson M et al. Strategy for Optimizing LC-MS Data Processing in Metabolomics: A Design of Experiments Approach. *Anal. Chem.*, 2012.

[40] Korkmaz, Goksuluk, and Zararsiz. MVN: An R Package for Assessing Multivariate Normality. *http://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf.*

[41] H. Lilliefors. On the Kolmogorov Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 1967.

[42] Moseley H.N.B. Error Analysis and Propagation in Metabolomics Data Analysis. *Comput Struct Biotechnol J.*, 2013.

[43] Gross J. and Ligges U. nortest: Five omnibus tests for testing the composite hypothesis of normality. *http://cran.r-project.org/web/packages/nortest/nortest.pdf.*

[44] Saeys Y., Inza I., , and Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007.

[45] de los Campos G. et al. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, 2013.

[46] Bartel J., Krumsiek J., and Theis F.J. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J.*, 2013.

[47] Wold S. Sjostrom M. and Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 2001.

[48] Hardoon D.R., Szedmak S., and Shawe-Taylor J. Canonical correlation analysis; An overview with application to learning methods. *Neural Computation*, 2004.

[49] Trygg J. and Wold S. Orthogonal Projections to Latent Structures (O-PLS). *J. Chemometrics*, 2002.

[50] Bylesj M. et al. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 2006.

[51] Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics*, 2002.

[52] Madsen R., Lundstedt T., and Trygg J. Chemometrics in metabolomics-A review in human disease diagnosis. *Analytica Chimica Acta*, 2010.

[53] Varmuza K. and Filzmoser P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. 2009.

[54] Tibshirani R. et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 2002.

[55] Sha W. et al. Metabolomic profiling can predict which humans will develop liver dysfunction when deprived of dietary choline. *FASEB*, 2010.

[56] Chen C. et al. Shrunken centroids regularized discriminant analysis as a promising strategy for metabolomics data exploration. *Journal of Chemometrics*, 2015.

[57] Krier C. et al. Feature clustering and mutual information for the selection of variables in spectral data. *European Symposium on Artificial Neural Networks*, 2007.

[58] Magendiran N. and Jayaranjani J. An Efficient Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data. *International Journal of Innovative Research in Science, Engineering, and Technology*, 2014.

[59] Srivastava D.K. and Bhambhu L. Data Classification Using Support Vector Machine. *Journal of Theoretical and Applied Information Technology*, 2005.

[60] Bhatia N. and Vandana A. Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security*, 2010.

[61] Boulesteix A. et al. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012.

[62] Kim S.B. et al. Controlling the False Discovery Rate for Feature Selection in High-resolution NMR Spectra. *Stat Anal Data Min.*, 2008.

[63] Fraser C.G. et al. Proposal for setting generally applicable quality goals solely based on biology. *Ann Clin Biochem*, 1997.

[64] Plebani M. et al. Performance criteria and quality indicators for the pre-analytical phase. *Clin Chem Lab Med*, 2015.

[65] Wald A. and Wolfowitz J. Tolerance limits for a normal distribution. *The Annals of Mathematical Statistics.*, 1946.

[66] Brown S.D., Ferr R.T.I., and Walczak B. *Comprehensive Chemometrics: Statistics, experimental design, optimization*. 2009.

[67] J. Prins. 7.2.6. What intervals contain a fixed percentage of the population values? In Croarkin C. and Tobias P., editors, *NIST/SEMATECH e-Handbook of Statistical Methods*. 2012.

[68] Young D.S. and Mathew T. Improved nonparametric tolerance intervals based on interpolated and extrapolated order statistics. *Journal of Nonparametric Statistics*, 2014.

[69] Muller P. and Quintana F. A. Nonparametric Bayesian Data Analysis. *Statistical Science*, 2004.

[70] Thirumuruganathan S. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. *https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/*.

[71] Bishop C. *Pattern Recognition and Machine Learning*. Springer, 2007.

[72] Cuperlovic-Culf M. *NMR Metabolomics in Cancer Research*. Woodhead Publishing, 2012.

[73] Watson G.S. Smooth Regression Analysis. *Sankhya: The Indian Journal of Statistics*, 1964.

[74] Nadaraya E.A. On Estimating Regression. *Theory Probab. Appl.*, 1964.

[75] Wold S. and Sjostrom M. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In Kowalski B.R., editor, *Chemometrics Theory and Application*. 1977.